SIMLIFE: Simulation of Long-term Human-Robot Interaction

Run Peng[1] Zinnia Nie[1] Jing Ding[1] Yinpei Dai[1] Yichi Zhang[2]
Zengqing Wu[3] Yao Fu[1] Ziqiao Ma[1,4] Jiayuan Mao[5] Joyce Chai[1]
[1]University of Michigan [2]ByteDance [3]Osaka University [4]Thinking Machines [5]Amazon
{roihn,chaijy}@umich.edu

Abstract

This is an abstract.

1. Introduction

Long-context understanding has become a core capability of modern foundation models, driving strong performance in applications such as conversational agents, coding assistants, and agentic systems. Extensive benchmarks have been developed to evaluate long-context memory and document understanding (e.g., SCROLLS [1], LongBench [2], L-Eval [3]), retrieval from extended inputs (e.g., Lost-in-the-Middle-style retrieval evaluations [4], RULER [5]), multi-hop or long-context reasoning (e.g., MuSiQue [6], BABILong [7]), and long-form multimodal/video understanding (e.g., EgoSchema [8], Video-MME [9], LVBench [10], EgoLife [11]).

For assistive agents that aim to provide lifelong companionship, it is crucial to maintain a personalized understanding of individuals shaped by rich, evolving temporal dynamics. EgoLife offers a pioneering step by extending from short, monographic videos to week-long observations, enabling the study of model behavior under ultra-long, real-life experiences. However, it primarily evaluates whether models can retrieve or summarize past events, rather than infer the latent mechanisms underlying human behavior. As a result, it falls short in assessing whether models can predict future behaviors or adapt to changes in routines and preferences.

To address this limitation, we propose SIMLIFE, a scalable simulated video benchmark that controls human behavior through deterministic rules.

2. Related Work

Long Video Understanding

Long-Context Understanding Models

3. SIMLIFE Dataset

3.1. Overview

SIMLIFE is a long-horizon multimodal dataset designed to study behavioral pattern understanding in situated environments. The dataset is generated using the life simulation game *The Sims 4*, where we simulate the daily activities of a household consisting of multiple family members (referred to as *Sims*) and a service robot named *Servo Bot*. This setup provides a controllable environment for generating realistic long-term behaviors and interactions.

We simulate 468 in-game days of household activities and record from the egocentric view of the Servo Bot. Each day covers the period from 6:00 AM to 10:00 PM and contains a variety of activities such as cooking, working, social interaction, and leisure. These simulated days serve as the fundamental building blocks of the dataset.

3.2. Data Generation Pipeline

The dataset is generated through a multi-stage simulation pipeline designed to produce long-term behavioral trajectories with controllable patterns. Each simulated character (Sim) is initialized with a behavioral profile specifying probabilistic preferences over activities such as cooking, fishing, fitness, etc. These profiles serve as the root distributions governing daily behavior.

Using these profiles, we sample activities to generate daily routines for the whole family. The resulting routines, which specify the sequence of activities for a day, are then compiled into scripts and executed in the game environment, yielding a pool of approximately 500 simulated days.

Next, we design behavioral patterns that describe recurring activity routines across multiple days. For example, a Sim may follow a pattern such as "Bob drinks coffee every weekday morning after staying up late." Given such a pattern, we derive a schedule that specifies what should or should not occur on each day of a sequence.

Finally, we construct video chains by selecting simulated days from the pool that satisfy the schedule requirements and linking them together in chronological order. Be-

cause of this modular design, individual daily videos can be reused across multiple video chains while still producing sequences that exhibit consistent behavioral patterns.

3.3. Data Modalities

SIMLIFE provides multiple synchronized modalities to support multimodal reasoning.

Video. Each simulated day is recorded as a 1080p video at 30 FPS using OBS[1]. A video spans approximately 24 minutes and corresponds to one in-game day from 6:00 AM to 10:00 PM. The camera follows the perspective of Servo Bot, providing an egocentric view of the household environment. As a result, the dataset naturally exhibits partial observability, reflecting the perceptual limitations of an embodied agent. The recordings also contain natural environmental audio generated by the game (e.g., footsteps, object interactions, and cooking sounds)[2].

Game Logs. In addition to visual observations, the dataset includes structured logs produced by the game engine. These logs record ground-truth actions and events occurring in the environment. For downstream use, we filter the logs to retain only events that are visible from the robot's perspective.

Dialogue and Speech Audio. Characters in the simulation may communicate with one another or interact with Servo Bot. Dialogue is generated conditioning on the scene context and behavioral profiles. Each Sim is assigned a unique voice, and speech volume dynamically adjusts based on spatial distance and obstacles such as walls, producing space-aware conversational audio. See more details in Appendix **??**.

Task-Oriented Dialogue. In addition to natural conversations, some video chains include task-oriented dialogue designed to embed information relevant to downstream tasks. These dialogues are inserted by controlling the speaker identities and conversation topics. When necessary, they replace the default generated dialogue in the corresponding segment of the video chain, enabling targeted evaluation scenarios such as retrieval or long-term reasoning.

4. Benchmark Design

4.1. Task Design

We propose a "pattern-in-the-haystack" task in a QA format to evaluate whether an agent can infer a simulated character's behavioral pattern from ultra-long observations. Each behavioral pattern consists of one or more main activities, supported by one or more deterministic rules conditioned on time or events.

For instance, the pattern "Bob drinks coffee every weekday morning or after staying up late" has the main activity "drink coffee" and includes two conditions: a temporal condition (every weekday morning) and an event condition (after staying up late). For each day in the context, we label whether these conditions are satisfied, which forms a truth table with different combinations of condition outcomes.

We then construct a video chain to simulate the pattern, ensuring that each combination of condition outcomes is observed at least three times. This provides sufficient evidence for an agent to form and verify hypotheses about the underlying behavioral pattern.

After all combinations have been observed at least three times, we present four types of questions. All questions are presented in a multiple-choice format, where the model selects the most likely statement or action from several candidates.

1. **Direct Prediction:** Predict the most likely behavior given the current conditions.
2. **Counterfactual Prediction:** Predict the most likely behavior under a hypothetical change in conditions.
3. **Inverse Reasoning:** Infer the most plausible condition or cause given a behavior.
4. **Noisy Counterfactual Prediction:** Predict the most likely behavior under hypothetical conditions that include irrelevant or distracting information.

After defining the four question types, we next determine when these questions are asked within the video chain. For a behavioral pattern with k conditions, there can be up to 2^k distinct combinations of condition outcomes. To thoroughly evaluate whether a model truly understands the underlying pattern, we probe the model under each of these possible combinations. For example, in the coffee-drinking pattern, questions may be asked under different situations such as weekday versus weekend and with or without staying up late the previous night. Each combination therefore defines a task instance, resulting in at most 2^k tasks for a video chain.

Each task consists of the four questions described above and is inserted immediately before the time when the main activity may occur. The correct answers depend on the condition outcomes of the selected day. A model must answer all four questions correctly to pass the task. Each question also includes carefully designed trap options that are semantically plausible but incorrect; they may correspond to unobserved behaviors, contradict the evidence, or be factually incorrect. This design discourages hallucination and shortcut reasoning, requiring models to consistently infer the underlying behavioral pattern from long-horizon observations. *[Roihn: We will have separate columns to show the*

[1]OBS Studio. https://obsproject.com/

[2]To comply with licensing constraints, background music and non-essential vocal effects are removed from the recordings.

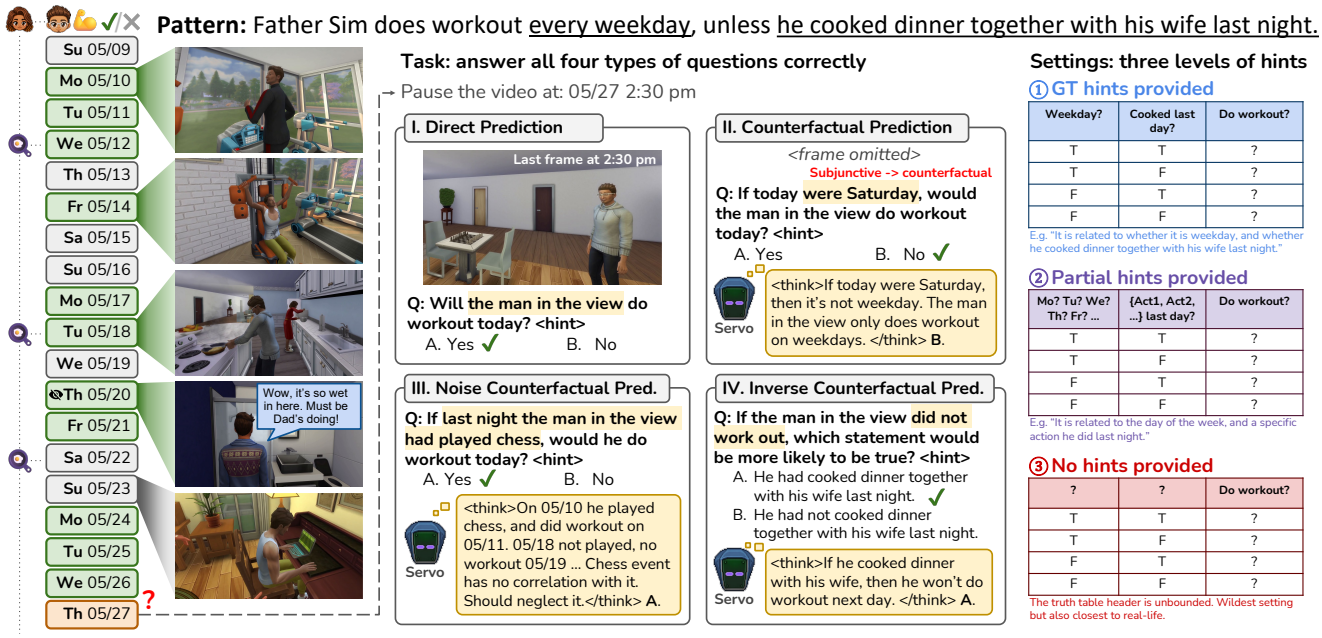

Figure 1. Example of SIMLIFE benchmark. Given the Father Sim's workout routine (shown at the top), we simulate three weeks of daily life that reflect this underlying pattern, recorded from the egocentric view of the Servo Bot. After sufficient observation, models are evaluated using four types of questions designed to assess their consistency in multi-source reasoning. We further provide three levels of hints that systematically ablate the skills required for pattern discovery, hypothesis elimination, and situated reasoning.

accuracy of each sub question in the main table, and then an overall success rate.]

4.2. Task Types

Pattern & Dynamic Understanding Our benchmark contains 222 video chains for *pattern understanding* and 56 for *dynamic understanding*. Pattern understanding tasks involve a single behavioral pattern and primarily evaluate a model's ability to infer a fixed behavioral rule from long-horizon observations. Dynamic understanding tasks contain multiple behavioral patterns and require models to identify different patterns as well as detect pattern shifts over time.

Intuitive & Counterintuitive Patterns To further stress the contextual reasoning required for question answering, we design half of the video chains (111 for pattern understanding and 28 for dynamic understanding) to follow *counterintuitive patterns*. Unlike intuitive patterns that align with common sense, counterintuitive patterns associate activities with unusual or non-plausible conditions. For example, a pattern such as *"drink coffee after playing the piano"* is unlikely to be inferred through common sense alone. Consequently, models must rely primarily on observations from the video context rather than prior knowledge when answering questions.

With & Without Predicates We introduce a task variant that controls whether the predicates defining a behavioral pattern are explicitly provided to the model. This variant allows us to distinguish between rule inference and rule discovery settings.

In the With Predicates setting, the conditions composing the pattern are given as part of the question context. The model therefore knows the relevant predicates and only needs to infer the behavioral rule that maps these predicates to the target activity. Since a pattern with k predicates induces 2^k possible combinations, and each combination is observed at least three times in the video chain, the model has sufficient evidence to identify the correct rule. A formal justification for the sufficiency of this observation design is provided in Appendix **??**.

In contrast, the Without Predicates setting does not reveal the predicates underlying the pattern. The model must first hypothesize which variables or events are relevant to the activity before inferring the rule itself. Because the hypothesis space is not bounded a priori, this setting is substantially more challenging and more closely resembles real-world pattern discovery. We provide a theoretical discussion on the validity of this formulation in Appendix **??**.

5. Result

6. Conclusion

References

[1] Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. SCROLLS: Standardized comparison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021, 2022. 1

[2] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023. 1

[3] Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-Eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*, 2023. 1

[4] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173, 2024. 1

[5] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024. 1

[6] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. 1

[7] Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. *arXiv preprint arXiv:2406.10149*, 2024. 1

[8] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, 2024. 1

[9] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Advances in Neural Information Processing Systems*, 2024. 1

[10] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. LVBench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024. 1

[11] Jiachen Yang et al. EgoLife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1